ПРАВИТЕЛЬСТВО РОССИЙСКОЙ ФЕДЕРАЦИИ
ФГАОУ ВО НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ УНИВЕРСИТЕТ
«ВЫСШАЯ ШКОЛА ЭКОНОМИКИ»

Факультет компьютерных наук
Образовательная программа «Программная инженерия»

УДК 004.62

Отчет об исследовательском проекте

на тему «Исследование методов обезличивания персональных данных для табличных данных»

(промежуточный, этап 1)

Выполнен студентом:
Группы #БПИ227 Шомполовым Максимом Андреевичем
ФИО студента

Проверен руководителем проекта:
Силаевым Юрием Владимировичем
ФИО, научная степень (если есть)
старшим преподавателем
Должность
Департамента программной инженерии
НИУ ВШЭ
Место работы (организация или департамент НИУ ВШЭ)

Москва 2024

РЕФЕРАТ

Важность проблемы обеспечения конфиденциальности персональных данных значительно выросла за последнее десятилетие. Одной из причин этого стало развитие интернета и цифровизация общества. Другая причина заключается в развитии анализа данных. Одним из способов решения данной проблемы является обезличивание персональных данных.

В настоящей работе проводится исследование методов и алгоритмов обезличивания табличных персональных данных. Предлагается сравнить качество подходов, описанных в различных источниках, с точки зрения риска раскрытия персональных данных и полезности обезличенных данных.

Для достижения данной цели будет произведён анализ различных методов обезличивания, описанных исследователями в данной области, а также различных подходов к оценке качества данных методов, реализация данных методов, подготовка данных для оценки качества работы алгоритмов обезличивания персональных данных, проведение экспериментов с реализованными алгоритмами и анализ результатов.

Данная работа состоит из 11 страниц, 1 таблицы, 1 приложения. Использовано 8 источников.

Ключевые слова: обезличивание, персональные данные, табличные данные, агрегация, пост-рандомизация, локальное подавление, перемешивание с обобщением, распределение данных, k-анонимность, l-разнообразие, t-близость.

СОДЕРЖАНИЕ

ТЕРМИНЫ И ОПРЕДЕЛЕНИЯ

Необезличенные персональные данные – персональные данные в виде, в котором они были получены от их владельца.

Персональная идентификационная информация (персональные идентификаторы) – любые данные, напрямую указывающие на их владельца.

Обезличенные персональные данные – данные, полученные в результате обработки необезличенных персональных данных, деобезличивание которых затруднено в результате данной обработки.

Деобезличиавние персональных данных – преобразование обезличенных персональных данных с целью целиком или частично установить соответствие между владельцем персональных данных и самими персональными данными.

Чувствительная информация – персональные данные не позволяющие идентифицировать их владельца, раскрытие которых несёт угрозу его анонимности.

Нечувствительная информация - персональные данные не позволяющие идентифицировать их владельца, раскрытие которых не несёт угрозу его анонимности.

Квазиидентификаторы – любые части персональных данных, которые позволяют определить, что какие-либо данные из данного набора не относятся к конкретному субъекту персональных данных.

Уровень анонимности обезличенных персональных данных – оценка рисков раскрытия владельца персональных данных.

Уровень полезности обезличенных персональных данных – свойство обезличенных персональных данных, позволяющее решать прикладные задачи с их использованием.

ВВЕДЕНИЕ

В настоящее время роль анализа данных значительно выросла. Он позволяет находить различные закономерности и проверять гипотезы в различных сферах: бизнесе, медицине, биологии, социальной сфере, экономике, финансах и других. Однако при обработке информации часто присутствует потребность работы с персональными данными. В этом случае возникает проблема получения согласия владельца персональных данных на их предоставление специалисту, который будет осуществлять анализ данных.

Другой проблемой, уже непосредственно в сфере защиты информации, является обеспечение безопасности информации, а именно её конфиденциальности, которая может быть нарушена в результате утечки данных. Актуальность этого вопроса в данный момент стала значительно выше из-за цифровизации общества.

Одним из подходов к решению данных проблем является обезличивание персональных данных. Оно представляет собой такую их обработку, что восстановление какого-либо соответствия между обработанными персональными данными и субъектом этих персональных данных затрудняется или становится невозможным. В результате данного преобразования часть информации теряется, но данные всё ещё в какой-то степени остаются содержательными для работы и анализа. При этом в зависимости от алгоритма обезличивания персональных данных степень сложности деобезличивания и степень полезности данных может меняться. Основные положения об обезличивании данных в России закреплены в Законе №152-ФЗ «О персональных данных» [1].

В настоящей работе проводится исследование различных методов обезличивания табличных персональных данных и их сравнения. Для этого проводится изучение самих алгоритмов обезличивания и подходов к оценке их качества с точки зрения рисков деобезличивания и полезности данных, а также реализация этих алгоритмов и подходов. На основе реализованных методов осуществляется сравнительный анализ способов обезличивания. Основными рассматриваемыми методами обезличивания являются агрегация, пост-рандомизация, локальное подавление, перемешивание с обобщением и распределение данных, а качество алгоритмов оценивается при помощи k-анонимности, l-разнообразия, t-близости.

1. ОБЗОР И АНАЛИЗ ИСТОЧНИКОВ

Рассмотрим источники, являющиеся отправной точкой при исследовании методов обезличивания табличных данных. Для понимания алгоритмов обезличивания требуется сначала рассмотреть способы оценки качества данных алгоритмов, так как подходы к обезличиванию в своей реализации опираются на критерии оценки рисков деобезличивания.

Первой статьёй, описывающей эти критерии, является [2]. В данной статье рассматриваются существовавшие на тот момент проблемы обезличивания данных, заключавшиеся в том, что уделение персональной идентификационной информации являлось недостаточным для обеспечения конфидициальности персональных данных. Для решения это проблемы был предложен декларативный подход к обезличиванию табличных данных и одновременно способ оценки его качества – k-анонимность. Также в статье рассматриваются конкретные способы обезличивания: обобщение и подавление. Также данный подход и его недостатки описываются в [3].

Как уже было сказано k-анонимность имеет некоторые недостатки, которые не позволяют ей обеспечивать достаточную конфиденциальность персональных данных в некоторых ситуациях. Часть из этих проблем решается при помощи подхода l-разнообразие, который является развитием k-анонимности и был предложен в источнике [4]. Однако данный метод также подвержен ряду рисков в случае, если обезличиваемые данные имеют определённую структуру. Эти недостатки описываются в статье [5]. Для их снижения в [5] предлагается использовать метод t-близости, который дополнительно учитывает распределение данных.

Теперь перейдём непосредственно к алгоритмам обезличивания. Основные алгоритмы представлены в [6], [7] и [8]. В [6] описываются основные положения обезличивания, в том числе критерии оценки качества обезличивания и некоторые алгоритмы обезличивания: псевдоанонимизация, подавление и частичное подавление, обобщение и агрегация, рандомизация, перестановка, взятие выборок. В [7] дополнительно рассматривается метод распределения данных. В [8] описывается применение методов обезличивания компанией Neoflex. Данные источники содержат ряд ссылок на статьи, в которых рассмотренные методы и алгоритмы описываются подробнее, в том числе раскрываются особенности вариантов их реализации.

2. АЛГОРИТИЫ, ВЫБРАННЫЕ ДЛЯ РЕШЕНИЯ ЗАДАЧ

Кратко опишем алгоритмы, которые будут применяться в данной работе для обезличивания персональных данных и оценки качества обезличивания.

2.2. Способы оценки качества алгоритмов обезличивания

2.2.1. k-анонимность

Формально данный метод можно описать следующим образом: говорят, что датасет является k-анонимным, если для каждой его записи существует ещё как минимум k-1 запись, такая что в каждой из них все квазиидентификаторы соответственно (по колонкам) совпадают. Таким образом, в случае идентификация конкретной записи по квазиидентификаторам конкретного человека становится невозможна, так как для каждого их набора будет существовать 0 или как минимум k соответствующих записей. Будем называть это множество записей корзиной.

Проблемой данного метода является возможное неравномерное распределение данных в датасете. В случае, если атрибут, содержащий чувствительную информацию, совпадает для всех k записей деобезличивание окажется возможным.

Данный метод требует значительно меньшей потери полезности информации, чем подходы, которые будут описаны ниже. Однако из-за этого риски раскрытия владельца персональных данных растут.

2.1.2. l-разнообразие

Определим данный подход формально. Датасет является l-разнообразным, если в каждой корзине из метода 2.1.1. для каждого признака, содержащего чувствительную информацию, существует хотя бы l различных значений. В таком случае, проблема, описанная в пункте 2.1.1. в некоторой степени решена, так как в k-записях или более с одинаковым набором квазиидентификаторов будут содержаться хотя бы l различных значений.

Несмотря на то, что при применении l-разнообразия становится невозможным точно определить чувствительную информацию для конкретного владельца персональных данных, всё ещё можно установить некоторое соответствие между обработанными персональными данными и субъектом этих персональных. Несмотря на то, что известно только то, что чувствительная информация владельца содержится в каком-либо множестве мощности хотя бы l, может оказаться, что этого достаточно, чтобы нанести ущерб владельцу персональных данных, в случае если каждый из l вариантов содержит в себе, например, дискредитирующую информацию.

Другим недостатком данного подхода является возможность того, что данные l значений будут неравномерно распределены в корзине или это распределение будет сильно

отличаться от распределения датасета. Например, в случае если k равняется 20, а l равняется 5, но при этом корзина содержит каждое из 4 значений только в одной записи, оставшееся значение в 16 записях, можно с уверенностью 80% утверждать, что запись субъекта персональных данных содержит часто встречающееся значение.

Данный метод обеспечивает более высокую защищённость от деобезличивания, но как показывают исследования, упоминающиеся в [5], даже при малых l происходит значительная потеря полезности данных.

2.1.3. t-близость

Снова сначала формально определим способ оценки. Датасет соответствует критерию t-близости, если расстояние между распределением данных в каждой колонке, содержащей чувствительную информацию, в каждой корзине из пункта 2.1.2. и распределением значений этой колонки во всём датасете на превышает t. При этом расстояние может рассчитываться при помощи различных подходов. Таким образом, t-близость решает проблемы l-разнообразия, описанные в пункте 2.1.2. Действительно в случае, если каждая корзина по своей структуре с точки зрения чувствительной информации похожа на весь датасет, то факт содержания записи пользователя в корзине не даёт больше информации, чем факт содержания записи пользователя в датасете (с точностью до t), что предполагается известным.

Основной проблемой этого метода является значительная потеря полезности данных при t, достаточных для решения проблем, описанных в пункте 2.1.2.

2.2. Алгоритмы обезличивания

2.2.1. Подавление и частичное подавление

Подавление заключается в исключении одного или нескольких атрибутов для каждой записи в датасете. Обязательным является применение подавления к персональным идентификаторам, так как они самостоятельно идентифицируют субъекта персональных данных.

Частичное подавление предполагает удаление отдельных атрибутов только из некоторых записей. Оно может быть применено к квазиидентификаторам, например, чтобы достичь k-анонимности.

Данные подходы могут приводить к некоторой потере полезности данных, однако искажения информации не происходит.

2.2.2. Обобщение и агрегация

Обобщение представляет собой замену значения на какой-либо диапазон. Например, замену возраста 36 лет на диапазон от 35 до 39 лет или замену даты 29 сентября 2024 года на сентябрь 2024 года.

Агрегация незначительно отличается от обобщения тем, что в результате весь диапазон разбивается на группы, например возраст от 20 до 24 лет, от 25 до 29 лет и так далее.

При применении данных подходов происходит искажение данных, но при этом полезность теряется значительно меньше, чем при использовании подавления.

2.2.3. Пост-рандомизация

При пост-рандомизации происходит добавление некоторой случайной величины к каждому значению какого-либо атрибута. От того, как распределена случайная величина, зависит степень искажения данных, а следовательно, степень их полезности и риск обезличивания.

2.2.4. Перемешивание с обобщением

Метод перемешивания с обобщением заключается в перестановке квазиидентификаторов между различными записями. При этом в случае, если предварительно было произведено обобщение переставляются группы значений. Это производится, например, для некоторого сохранения полезности данных. В результате искажения данных, полученных после обобщения, не происходит.

2.2.5. Распределение данных

Подход заключается в том, чтобы не предоставлять пользователю всех данных одновременно, но при этом оставлять возможность выполнять аналитические запросы. Это реализуется за счёт использования забывчивой передачи информации, которая является протоколом с нулевым разглашением. При передаче данных с использованием классического алгоритма отправитель формирует два набора данных, а получатель выбирает, какой из наборов он хочет получить. При этом отправитель не имеет возможности узнать, что выбрал получатель, а получатель не может узнать все данные целиком.

Исследователями в данной области были также разработаны другие подходы к обезличиванию табличных персональных данных, такие как псевдоанонимизация, однако в данном разделе описаны методы, использующиеся в настоящей работе.

СПИСОК ИСПОЛЬЗОВАНЫХ ИСТОЧНИКОВ

1. Федеральный закон от 27.07.2006 №152-ФЗ «О персональных данных»
2. Samarati P., Sweeney L. Protecting privacy when disclosing information: k-anonymity and its enforcement through generalization and suppression. – 1998
3. k-anonymity, the parent of all privacy definitions // https://desfontain.es URL: https://desfontain.es/blog/k-anonymity.html (дата обращения: 01.12.2024).
4. Machanavajjhala A. et al. l-diversity: Privacy beyond k-anonymity //Acm transactions on knowledge discovery from data (tkdd). – 2007. – Т. 1. – №. 1. – С. 3-es.
5. l-diversity, because reidentification doesn't tell the whole story // https://desfontain.es URL: https://desfontain.es/blog/l-diversity.html (дата обращения: 01.12.2024).
6. Garfinkel S. L. De-Identification of Personal Information //National Institute of Standard and Technology. – 2015.
7. Borisov S. A., Bosov A. A., Ivanov D. E. APPLICATION OF SIMULATED COMPUTER SIMULATION TO THE TASK OF PERSONAL DEPERSONALIZATION DATA. MODEL AND ALGORITHM FOR DECONTAMINATION BY SYNTHESIS //Programmirovanie. – 2023. – №. 5. – С. 19-34.
8. Как маскировка данных спасает вашу приватность // habr URL: https://habr.com/ru/companies/neoflex/articles/820333/ (дата обращения: 01.12.2024).

ПРИЛОЖЕНИЕ 1

КАЛЕНДАРНЫЙ ПЛАН РАБОТ

Таблица 1 – Календарный план работ

Этап проекта	Описание работ	Ожидаемые результаты	Сроки выполнения
1. Анализ методов обезличивания	Рассмотреть описанные исследователями алгоритмы обезличивания и особенности их реализации. Выбрать алгоритмы для последующей реализации.	Наличие перечня алгоритмов для реализации и знание особенностей их работы.	06.11.2024 – 25.01.2025
2. Анализ подходов к оценке качества обезличивания	Рассмотреть описанные исследователями алгоритмы оценки качества обезличивания и особенности их реализации. Выбрать алгоритмы для последующей реализации.	Наличие перечня алгоритмов для реализации и знание особенностей их работы.	06.11.2024 – 25.01.2025
3. Подготовка данных	Поиск готовых данных и (или) подготовка своего датасета.	Наличие нескольких готовых датасетов, содержащих табличные персональные данные различных типов.	03.01.2024 – 20.01.2024
4. Реализация методов оценки и обезличивания	Реализация методов, выбранных на этапах 1 и 2 на языке программирования Python.	Реализованы описанные алгоритмы.	25.12.2024 – 10.03.2025
5. Проведение экспериментов	Проведение экспериментов с реализованными алгоритмами и подготовленными данными.	Получены результаты экспериментов, позволяющие сравнить алгоритмы между собой.	10.01.2025 – 25.03.2025
6. Анализ результатов и написание отчёта	Проведение сравнения качества обезличивания при помощи реализованных алгоритмов и написание отчёта.	Подготовлен отчёт об исследовательском проекте.	30.01.2025 – 15.04.2025